DOI:10.19651/j.cnki.emt.2211181

基于多智能体强化学习值分解的优化算法*

王慧琴 苗国英 孙英博
(南京信息工程大学自动化学院 南京 210044)

摘 要: 当前多智能体强化学习在值分解的算法中无法充分考虑到多智能体间的协作关系,并且使用的随机策略在探索过程中容易出现越过最优点,陷入局部最优解的情况。针对以上问题,本文提出了一种深度交流多智能体强化学习算法。本文通过使用卷积和全连接结构在值分解网络中设计了一种通信机制以此来增强多智能体之间的协作。接着,本文提出了一种新的自适应探索策略,为了平衡数据探索与利用之间的矛盾,加入了周期性的衰减策略。最后,通过仿真结果验证了本文提出方法在部分场景中达到 25.8%的性能提升,提高了多智能体的合作能力。

关键词: 多智能体;强化学习;深度交流;探索策略
中图分类号: TP242 **文献标识码:** A **国家标准学科分类代码:** 520.2050

Optimization algorithm based on value decomposition of multi-agent reinforcement learning

Wang Huiqin Miao Guoying Sun Yingbo
(School of Automation, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: The current multi-agent reinforcement learning algorithm cannot fully consider the cooperative relationship between multi-agents in the value decomposition algorithm, and the stochastic strategy used in the exploration process is prone to cross the optimal point and fall into the local optimal solution. Aiming at the above problems, this paper proposes a deep communication multi-agent reinforcement learning algorithm. This paper designs a communication mechanism in value decomposition network by using convolution and fully connected structure to enhance the cooperation between multi-agents. Then, a new adaptive exploration strategy is proposed in this paper. In order to balance the contradiction between data exploration and utilization, a periodic decay strategy is added. Finally, simulation results verify that the proposed method achieves 25.8% performance improvement in some scenarios, and improves the cooperation capability of multi-agent.

Keywords: multi-agent;reinforcement learning;deep communication;explore the strategy

0 引 言

近年来,多智能体强化学习(multi-agent reinforcement learning,MARL)在自动驾驶、路径规划、编队控制、机器人协作等领域[1]有大量研究,越来越受到研究者的关注。此外,随着计算和存储能力的大幅提升,强化学习[2]在人工智能领域获得了巨大成功[3]。强化学习将感知、学习、决策融合到同一框架,实现了从原始输入到决策动作"端到端"的感知与决策,并在游戏领域取得了优异的成绩。Google DeepMind 公司研发的 AlphaGo 系列围棋程序,击败了顶尖职业选手[4-6];提出的深度 Q 网络(deep Q-network,DQN),在多种 Atari 游戏中成功超越人类专业玩家。OpenAI 研发了能够在 Dota2 这一比围棋更复杂的游戏中击败人类专业玩家的游戏机器人[7]。

然而,强化学习运用到多智能体领域具有重大挑战[8-9]:1)智能体数量上的增多会导致状态空间和动作空间呈指数增长,造成空间维度爆炸;2)单个智能体的探索会对其他智能体的决策产生影响,从而引起环境的不稳定;3)每个智能体并不能得到全部环境信息,而只能对部分环境信息进行观测。

为了解决上述问题,研究者们提出大量算法:其中,Foerster 等[10]利用 actor-critic 框架提出的 COMA 算法(counterfactual multi-agent,COMA)不受环境的不平稳性影响,其核心之处在于引入了一个反事实的基线函数,通过

收稿日期:2022-08-25
* 基金项目:国家自然科学基金(62073169)、江苏省"333 工程"项目(BRA2020067)资助

使用联合的 critic 去计算每个智能体独自的优势函数，遵循当前 actor 网络进行决策得到的全局回报与反事实基线之间的差值，求得该智能体的 Q 值，但其可扩展性较差。然后，Sunehag 等[11]和 Rashid 等[12]在值分解基础上分别提出的 VDN 算法（value-decomposition networks，VDN）和 QMIX 算法（monotonic value function factorisation，QMIX），依赖于简单的结构约束，VDN 采用的方法就是直接相加求和，对每个智能体的值函数进行整合，得到一个联合动作值函数 $Q_{tot}(\tau,u)$，单个智能体 i 的局部动作值函数为 $Q_i(\tau^i,u^i)$，全局值函数依赖于单个智能体的值函数，QMIX 是在联合动作值函数和单个智能体动作值函数上增加了一种单调性约束，接着采用混合网络对每个单智能体的动作值函数做非线性映射，并保证其权重非负。但是 VDN 和 QMIX 存在结构上的约束，只能解决部分简单的多智能体任务。因此，Son 等[13]提出一种放宽约束的算法 QTRAN（learning to factorize with transformation，QTRAN）直接学习 Q_{total}，同时创造了两个条件来约束 Q_{total} 和 Q_i 之间的关系，从而更新 Q_i。考虑到 QTRAN 的结构设计复杂，计算成本较高且低效探索，Mahajan 等[14]提出 MAVEN 算法（multi-agent variational exploration，MAVEN），MAVEN 将值分解与策略的方法混合，其中基于值的智能体根据分层控制策略的共享潜在变量来约束其行为，更好的解决多智能体的任务。上述成果发现，虽然值函数分解方法在一定程度上提高了多智能体的学习效率，但在同一环境下存在收敛性缓慢且不稳定，其变化主要受环境的非完全观测、决策探索的低效性等因素影响。

本文针对上述问题，提出了一种促进多智能体之间信息交流，增强策略的算法，称为深度交流多智能体强化学习（deep communication multi-agent reinforcement learning，DCMR）算法。通过引入由特征提取模块（feature extraction module，FEM）和权重提取模块（weight extraction module，WEM）构成的通信机制用来过滤信息量，使智能体之间的通信便变得高效简洁，避免造成资源浪费。此外，为了提高模型的鲁棒性，对传统的策略进行改进，设计周期性探索策略。仿真结果证明本文所提方法的有效性。

1　深度强化学习理论

1.1　深度 Q 网络

DQN 中有两个结构完全相同但是参数却不同的网络[15-16]，一个用于预测 Q 估计（eval_net），使用最新的参数，一个用于预测 Q 现实（target_net），使用未更新的参数，Q 现实的计算为：

$$Q_{tar}=r+g\times Q_{\max}(s',a',q) \tag{1}$$

根据 Q 现实与 Q 估计得到损失，损失函数一般采用均方误差损失

$$LOSS(\theta)=E[(Q_{tar}-Q(s,a;\theta))^2] \tag{2}$$

此外，DQN 引入了经验池[17]，在学习过程中，DQN 智能体的经验 (s_t,a_t,r_t,s_{t+1})，其中 s_t、a_t 和 r_t 分别表示在时间步长 t 时的状态所选动作和收到的奖励，s_{t+1} 表示下一个时间步长的状态。为了更新 Q 值，它使用随机小批量更新，在训练时从经验回放存储器中均匀随机采样。

1.2　马尔可夫决策过程

马尔可夫决策过程（Markov decision process，MDP）[18]由 $<S,A,\boldsymbol{P},r,\gamma>$ 构成，其中 S 是有限状态的集合，$\boldsymbol{P}$ 是状态转移矩阵，$r(s,a)$ 是奖励函数，$P(a\mid s)$ 是状态转移函数，表示在状态 s 执行动作 a 之后到达 s' 状态的概率。马尔可夫决策过程与马尔可夫奖励过程相比存在一个智能体来执行动作，智能体根据当前状态从动作的集合 A 中选择一个动作的函数，称为策略。$\pi(a\mid s)=P(A=a\mid S=s)$，表示在输入状态 s 情况下采取动作 a 的概率。强化学习的目的就是寻求最优策略 $\pi(a\mid s)$ 使得智能体获取到的累计奖励最大化[19]。

1.3　卷积神经网络

由输入层、卷积层、激活函数、池化、全连接层组成[20]，其中卷积层用来特征提取和权值共享。全连接层连接所有的特征，将输出值送给分类器。卷积神经网络中前几层的卷积层参数量占比小，计算量占比大；后面的全连接层正好相反，大部分卷积神经网络都具有这个特点。

2　DCMR 算法

这一节中，利用深度强化学习提出改进策略，并增强多智能体间的信息交流，称为 DCMR 算法。DCMR 算法主要在 QMIX 算法框架的基础上进行改进。DCMR 算法的整体框架如图 1 所示。

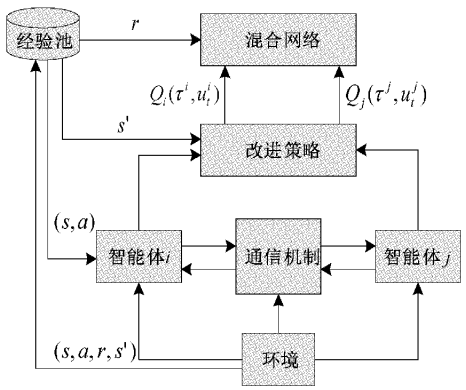


图 1　DCMR 算法框架

2.1　通信机制

由于多智能体系统的目标是将多智能体间大而复杂的系统转化为小而相互协调的系统，因此本文在分布式执行的框架上，利用深度学习技术来学习提取输入数据的抽象特征，并优化智能体在强化学习问题中的决策。所提方法可以应用在以值分解为基础的强化学习算法上，每个智能体基于联合动作值的最优联合动作与基于个体动作值的个

体动作的联合达到一致为目的,即单个全局最大值(individual global max,IGM)原则。

1) 网络结构

与现有的 MARL 中多智能体间通信的结构不同,例如 CommNet[21] (multi-agent communication with backpropagation)复杂度较高,容易维度爆炸、BiCNet[22] (multi-agent bidirectionally-coordinated nets)专门应用于特定设计和 ATOC[23](attentional communication model)泛化性能有限。结合这些缺点,本文提出的通信机制在智能体开始交互环境时,就采取措施对信息进行筛选,提高利用效率,从而减少一些无谓的资源浪费,整体结构如图 2 所示。在主支路中首先对输入序列进行维度的调整,假设输入序列的长度为 C,经过维度的调整后则增加了一个维度,接着利用 FEM 来提取信息特征,同时添加一条辅助支路对输入数据进行权重计算,最后得到的权重与特征信息进行乘积,以此来关注输入数据中的重要信息,过滤无用的信息,避免资源浪费,提高效率。

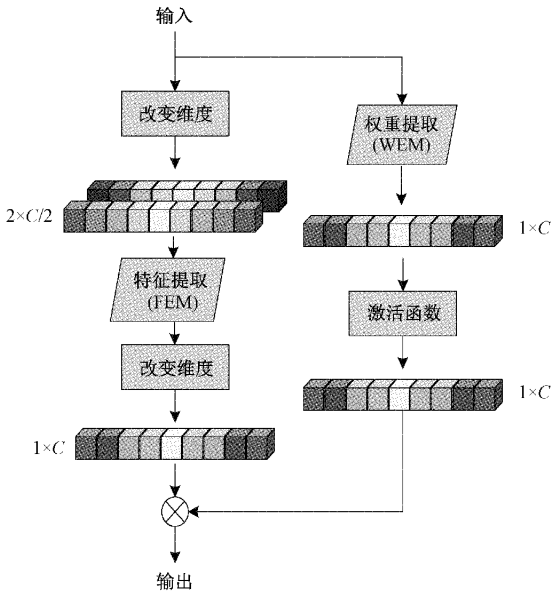


图 2 通信机制网络

该网络中的输入分别经过 FEM 和 WEM 模块,将大量特征进行各种变换后传出。WEM 模块和 FEM 模块的结构分别如图 3 所示。

2) 全连接层

WEM 主要由全连接层构成,如图 3(a)所示其一共包含 3 层:Layer1, Layer2, Layer3,每一层由线性函数(Linear)、归一化函数(LayerNorm)和激活函数(ReLU)组成,为了减少信息的冗余,避免过拟合,在层数之间加入丢弃层(Dropout),其中设定的每层的丢弃率为 0.2。

3) 卷积层

如图 3(b)所示,该模块主要用来强化特征信息的提取,主要由多个卷积层构成。将调整维度之后的输入序列

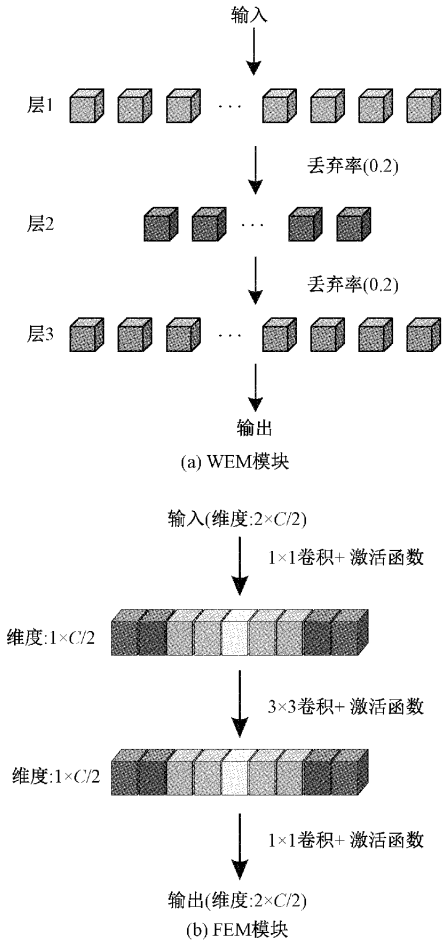


图 3 模块结构

进行输入,一共经过 3 层卷积层,每层由卷积层和激活函数构成。模块首先经过一个 1×1 的卷积层来是实现跨通道信息之间的交互,接着使用 3×3 卷积核对特征信息进行提取,最后再次利用 1×1 卷积层实现对特征信息的调整筛选。另一方面,对于这样一个改进的通信方法也会对之后训练以及最终决策产生影响。总的来说,通信机制对智能体的沟通更加有效。

2.2 改进策略

贪心策略是强化学习中一种常用的探索策略(ε-greedy)。每个智能体网络结构采用 MLP 和 GRU 组成的深度循环 Q 网络,按照式(3)所示。

$$\pi(a\mid s)=\begin{cases}1-\varepsilon_t+\dfrac{\varepsilon_t}{|A|}, & a=A^*\\ \dfrac{\varepsilon_t}{|A|}, & a\neq A^*\end{cases} \tag{3}$$

选取动作,有 ε 的概率会随机选择一个动作,$1-\varepsilon$ 的概率从已有动作中选择价值最大的动作。其中,A^* 表示最优动作,$|A|$ 表示当前智能体状态所有可选动作的集合,为动作观测历史。

在有限马尔科夫决策过程中,虽然每个动作都有被选

择的概率，但是这种选择太过于随机，一直是以 ε 的概率进行探索，可执行或不可执行的动作都会被相同的概率探索，这就导致智能体收敛速度过慢，同时不能更好的获得最优动作。因此，本文提出一种周期性衰减探索(exploration of periodic decay，EP-ε)，定义此探索表示为：

$$\varepsilon_t = e^{-\frac{t}{a}} \lg[9 \times |\cos\frac{2(T-t)}{T}\pi| + 1] \tag{4}$$

该方法与文献[24]提到的自适应搜索策略不同，并结合余弦退火[25-26]的思想进行改进。其中，$a = 200\ 000$，T 为周期，$t \in [0, N)$；T 越大，探索越缓慢。由于智能体刚开始学习时，不能真正学习到知识，所以前期给定较大的探索值，接着在邻近的一个极小点位置时探索率开始减小，使其尽可能靠近这个极小点，然后再加速到达下一个这样的位置附近时再减小，这样以一种周期性缓慢的形式衰减，衰减过程如图 4 所示。

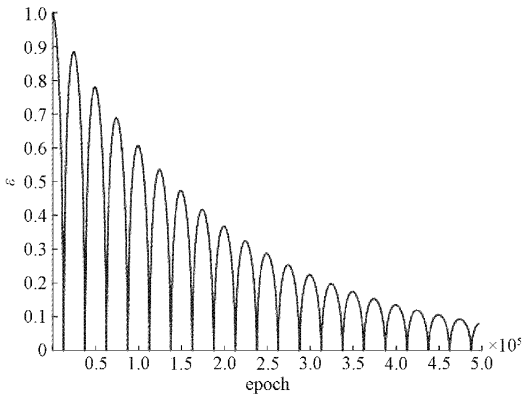


图 4　探索衰减过程

根据上述思想，贪婪率的选择避免了传统方式的以下缺点，1)避免了在贪婪率过大的时候，整个搜索速度会很快，导致直接越过最优点；2)避免了在其更小的时候，虽然可以找到最佳解，但若一直保持不变，在到达其中一个最优点的时候，模型会自动认为已经达到全局最优点，便再难找出全局理想值。

2.3　训练过程

步骤 1)每个智能体在与环境交互时，首先从环境中获得的观测量 O_n 作为输入，通过多层全连接层对信息的筛选后经过卷积变换提取出当前数据中的重要信息的权重，接着将提取到的权重信息与原始数据进行乘积，将权重进行分配，关注其中有意义的信息 C_i。

步骤 2)经过通信机制的多智能体，输出的 $Q(\tau, c, u)$ 作为改进策略的输入，使得可以更好地平衡探索和利用策略，再根据式(3)和(4)，输出独立的值函数 $Q(\tau, u)$，然后输入混合网络。

步骤 3)接受每个智能体的输出 Q 值，以及从经验回放里随机采样 b 个样本数据经过混合网络，对每个子智能体的动作值函数做非线性映射，得到 Q 最大值的总和。

3　仿真实验

3.1　无人车群智能交通

本实验全部都是基于 PyTorch 深度学习框架，版本号为 1.10.0，Python 版本号为 3.8.12。硬件环境上使用英伟达系列显卡来进行训练，显卡型号为 NVIDIA GeForce RTX 3080，显存为 10 G；CPU 型号为 AMD 系列 R5-5600x，内存为 16 G。

针对无人车群在一定区域内遇到道路上的窄道和其他无人车的问题，本实验在 4×9 的网格上进行仿真训练。如图 5 所示，本例中共有 4 个智能体(无人车)，它们的目标是在不碰到其他智能体和超出路线的前提下互相协作，依次驶过中间狭窄的车道到达对面，每个网格中在同一时间最多只能出现一个无人车。若撞到边界或其他无人车，则判定此次动作是无效的，并给予相应的惩罚。为了保证无人车在到达对面时做出的每一步动作都是有效的，当无人车进入已经驶过的区域时，也判定此次动作是无效的，同样会给予相应的惩罚。

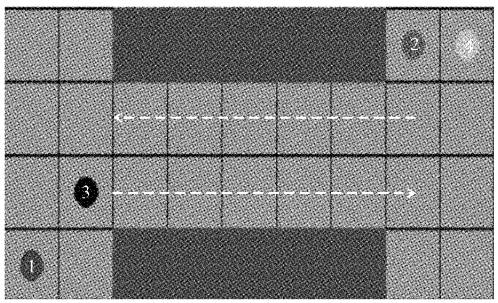


图 5　二维环境示意图

2.1 节中提出的通信机制有很好的协作效果，为测试改进后算法的收敛性，在基于相同参数条件的基础上，经过 6 000 次的迭代训练，其奖励回报情况如图 6 所示。从图中可以看出来，与原始方法 QMIX 相比较，本文改进后的方法在后期逐渐平稳，碰撞次数减少，性能更优。

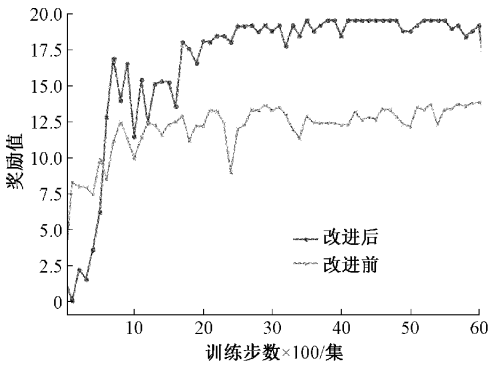


图 6　累积奖励值

3.2　预演战争模拟仿真

在一定规则下计算战场状态的重要手段便是兵棋推演，随着人工智能与兵棋推演结合的兴起，备受各军事强国关注。并且强化学习作为人工智能最热门的方法之一，应

用于兵棋推演环境最为合适。由于兵棋推演与即时战略游戏都考虑到战役战术的层面,因此本文选取星际争霸微观操作环境 SMAC 作为预演战争的仿真平台。该环境中,智能体(兵棋)的动作空间由一系列的离散动作组成:移动方向(东南西北)、攻击敌人(仅当敌人出现在涉及范围内)、停止和空闲(仅当阵亡时)。且每个智能体都是独立的,这样的智能体形成群体与内置的脚本 AI 对战,对战中目的是使智能体(兵棋)学会自己探索,找出最合适的策略获得胜利。

1) 场景描述

该实验在《星际争霸Ⅱ》[27]环境中进行,让玩家能够学习协作智能体(兵棋)之间的复杂互动。表 1 中的实验场景是为了验证 DCMR 算法在多种情形下的训练学习能力。

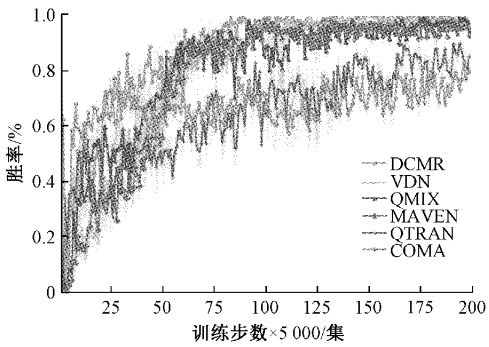
表 1 实验场景

场景名称	敌我数量	场景类型
3m_Map	3(友)vs3(敌)	同构 & 对称
2s3z_Map	5(友)vs5(敌)	异构 & 对称
3s5z_Map	8(友)vs8(敌)	异构 & 对称

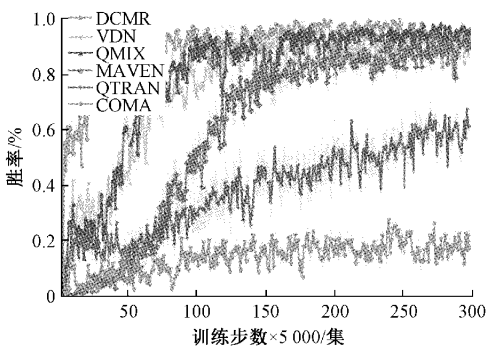
注:m 代表 Marines,s 代表 Stalkers,z 代表 Zealots

2) 实验结果与分析

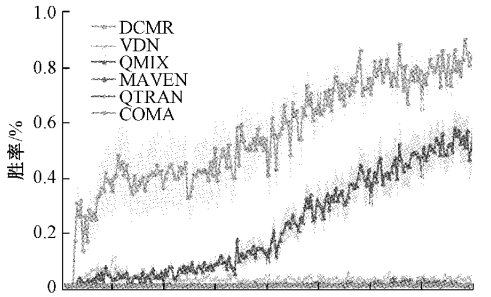
将 DCMR 算法与其他 5 种算法(VDN,QMIX,MAVEN,QTRAN,COMA)进行对比。在 3m_Map 和 3s5z_Map 地图训练 100 万个时间步数,2s3z_Map 地图训练 150 万个时间步数。在各实验地图独立实验 8 次,结果如图 7(a)~(c)所示。



(a) 3 m地图



(b) 2s3z地图



(c) 3s5z地图

图 7 模拟对抗实验结果

图 7 中实线与阴影分别代表 8 次实验的均值与浮动幅度。实验每进行 5 000 次 episode,保存一次模型,经验复用池容量为 5 000 条经验量。

通过表 2~4 展示了 6 种算法在以上 3 类地图训练的结果,数据均以实线为准。表中性能栏第 1 列、第 2 列和第 3 列均代表训练到一定时间步数的胜率,时间步长数据均为训练步数 × 5 000,在其中两个地图上的最高胜率为 100%。

表 2 3 m 地图训练胜率

算法对比	75	150	200
DCMR	94.60	96.9	95.7
VDN	88.10	96.8	95.5
QMIX	77.20	92.5	94.3
MAVEN	85.00	94.3	92.4
QTRAN	61.80	80.0	91.9
COMA	67.90	70.7	79.3

表 3 2s3z 地图训练胜率

算法对比	100	200	300
DCMR	95.7	92.7	95.6
VDN	71.5	88.9	89.1
QMIX	86.9	91.4	95.1
MAVEN	47.7	82.2	89.4
QTRAN	25.3	46.3	64.5
COMA	14.5	14.3	22.2

表 4 3s5z 地图训练胜率

算法对比	75	150	200
DCMR	39.3	70.7	85.50
VDN	3.2	2.4	51.74
QMIX	15.0	36.8	59.70
MAVEN	0.4	1.0	2.60
QTRAN	0	0.3	0.80
COMA	0	0	0

从以上图表结果分析可知，在 3m_Map 上的多智能体可以非常好的学习到合适的策略，6 种算法都有上升，其中有四种算法更加趋于收敛，胜率相差不大。在 2s3z_Map 上仍然是这 4 种算法在不断收敛，不同的是它们的收敛速度相距较大。在 3s5z_Map 上只有两个算法在不断学习，其余 4 个算法效果不明显或者就是没有学会。根据结果显示，本文的 DCMR 算法提出的通信机制以及策略改进才能使得智能体更好的互相协作，收敛更快。可以进一步在兵棋推演智能决策中运用，为人工智能和兵棋推演的结合提供参考。

4　结　　论

对于协作式多智能体系统而言，提出一种增强多智能体合作能力的 DCMR 算法，使多智能体在非完全观测的环境下效果突出。改进了多智能体的探索能力，促进其探索的平衡力。

在多个实验场景下展示 DCMR 算法确实具有巨大优势，在一定程度上提升了效率，加快了收敛的速度性。

未来工作中，随着智能体数目的增加，系统的复杂性将不断提高，DCMR 算法还需优化卷积神经网络的结构来提高处理速度，减少计算量。由于卷积只关注到局部特征之间的关系，无法建立长距离信息之间的联系，所以距离较远的智能体之间的信息关联得不到充分的关注。如何构建高效、通用性更强的神经网络来不断拟合和完善实际中的应用也是需要解决的问题。

参考文献

[1] 孙彧，曹雷，陈希亮，等. 多智能体深度强化学习研究综述[J]. 计算机工程与应用，2020，56(5)：13-24.

[2] SUTTON R S，BARTO A G. Introduction to reinforcement learning [M]. Cambridge：MIT Press，1998.

[3] 孙长银，穆朝絮. 多智能体深度强化学习的若干关键科学问题[J]. 自动化学报，2020，46(7)：1301-1312.

[4] 周志华. AlphaGo 专题介绍[J]. 自动化学报，2016，42(5)：670.

[5] SILVER D，HUANG A，MADDISON C J，et al. Mastering the game of go with deep neural networks and tree search [J]. Nature，2016，529 (7587)：484-489.

[6] SILVER D，SCHRITTWIESER J，SIMONYAN K，et al. Mastering the game of go without human knowledge[J]. Nature，2017，550(7676)：354-359.

[7] BERNER C，BROCKMAN G，CHAN B，et al. Dota 2 with large scale deep reinforcement learning [J]. ArXiv Preprint，2019，ArXiv：1912.06680.

[8] NGUYEN T T，NGUYEN N D，NAHAVANDI S. Deep reinforcement learning for multiagent systems：A review of challenges，solutions，and applications[J]. IEEE transactions on cybernetics，2020，50 (9)：3826-3839.

[9] 熊丽琴，曹雷，赖俊，等. 基于值分解的多智能体深度强化学习综述[J]. 计算机科学，2022，49 (9)：172-182.

[10] FOERSTER J，FARQUHAR G，AFOURAS T，et al. Counterfactual multi-agent policy gradients [C]. Proceedings of the AAAI Conference on Artificial Intelligence，2018：2974-2982.

[11] SUNEHAG P，LEVER G，GRUSLYS A，et al. Value-decomposition networks for cooperative multi-agent learning based on team reward[C]. Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems，2018：2085-2087.

[12] RASHID T，SAMVELYAN M，SCHROEDER C，et al. QMIX：Monotonic value function factorisation for deep multi-agent reinforcement learning [C]. International Conference on Machine Learning，2018：4295-4304.

[13] SON K，KIM D，KANG W J，et al. QTRAN：Learning to factorize with transformation for cooperative multi-agent reinforcement learning [C]. International Conference on Machine Learning，2019：5887-5896.

[14] MAHAJAN A，RASHID T，SAMVELYAN M，et al. MAVEN：Multi-agent variational exploration [C]. Advances in Neural Information Processing Systems，2019：7611-7622.

[15] 许杨子，强文，刘俊，等. 基于改进深度强化学习算法的电力市场监测模型研究[J]. 国外电子测量技术，2020，39(1)：82-87.

[16] 周震尘，金涛. 一种极端自然事件下的基于深度强化学习的配电网脆弱性研究方法[J]. 中国测试，2022，48(2)：98-104.

[17] ZHA D，LAI K H，ZHOU K，et al. Experience replay optimization [J]. ArXiv Preprint，2019，ArXiv：1906.08387.

[18] WHITE C. Markov decision processes [M]. New York：Springer，2001.

[19] 康守强，刘哲，王玉静，等. 基于改进 DQN 网络的滚动轴承故障诊断方法[J]. 仪器仪表学报，2021，42(3)：201-212.

[20] 毛向向，王红军，韩凤霞，等. 基于深度卷积神经网络的机电系统故障分类识别方法[J]. 电子测量与仪器学报，2021，35(2)：87-93.

[21] SUKHBAATAR S，FERGUS R. Learning multiagent communication with backpropagation[C]. Advances in

Neural Information Processing Systems, 2016: 2244-2252.

[22] PENG P, WEN Y, YANG Y, et al. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starCraft combat games[J]. ArXiv Preprint, 2017, ArXiv: 1703.10069.

[23] JIANG J, LU Z. Learning attentional communication for multi-agent cooperation[C]. Advances in Neural Information Processing Systems, 2018: 7254-7264.

[24] 张强,李盼池. 一种自适应多策略行为粒子群优化算法[J]. 控制与决策, 2020, 35(1): 115-122.

[25] GAO H, LI Y, PLEISS G, et al. Snapshot Ensembles: Train 1, get M for free[J]. ArXiv Preprint, 2017, ArXiv: 1704.00109.

[26] LOSHCHILOV I, HUTTER F. SGDR: Stochastic gradient descent with warm restarts[C]. ICLR 2017 5th International Conference on Learning Representations, 2016.

[27] RICHOUX F. Terrain analysis in starcraft 1 and 2 as combinatorial optimization[J]. ArXiv Preprint, 2022, ArXiv: 2205.08683.

作者简介

王慧琴,硕士研究生,主要研究方向为多智能体强化学习。

E-mail:1229066030@qq.com

苗国英,副教授,硕士生导师,主要研究方向为多智能体系统协调控制。

E-mail:mgyss66@163.com

孙英博,硕士研究生,主要研究方向为多智能体强化学习。

E-mail:sun02070207@163.com